doi:10.1006/geno.2001.6655, available at http://www.idealibrary.com on IDEAL

Assembly and Annotation of Human Chromosome 2q33 Sequence Containing the *CD28*, *CTLA4*, and *ICOS* Gene Cluster: Analysis by Computational, Comparative, and Microarray Approaches

Vincent Ling,* Paul W. Wu, Heather F. Finnerty, Michael J. Agostino, James R. Graham, Sanjun Chen, Jason M. Jussiff, Gregory J. Fisk, Christopher P. Miller, and Mary Collins

Genetics Institute/Wyeth Research, 87 Cambridge Park Drive, Cambridge, Massachusetts 02140, USA

To whom correspondence and reprint requests should be addressed. Fax: (617) 665-5584. E-mail: VLing@genetics.com.

Human chromosome 2q33 is an immunologically important region based on the linkage of numerous autoimmune diseases to the *CTLA4* locus. Here, we sequenced and assembled 2q33 bacterial artificial chromosome (BAC) clones, resulting in 381,403 bp of contiguous sequence containing genes encoding a NADH:ubiquinone oxidoreductase, the costimulatory receptors CD28, CTLA4, and ICOS, and a HERV-H type endogenous retrovirus located 366 bp downstream of *ICOS* in the reverse orientation. Genomic microarray expression analysis using differentially activated T-cell RNA against a subcloned *CTLA4/ICOS* BAC library revealed upregulation of *CTLA4* and *ICOS* sequences, plus antisense *ICOS* transcripts generated by the HERV-H, suggesting a potential mechanism for *ICOS* regulation. We identified four nonlinked, polymorphic, simple repetitive sequence elements in this region, which may be used to delineate genetic effects of *ICOS* and *CTLA4* in disease populations. Comparative genomic analysis of mouse genomic *Icos* sequences revealed 60% sequence identity in the 5′ UTR and regions between exon 2 and the 3′ UTR, suggesting the importance of ICOS gene function.

Key words: CD28, CTLA4, ICOS, CD152, autoimmunity, genetics, microarray, diabetes, microsatellite, retrovirus

INTRODUCTION

Immunological activation of T cells is accomplished by a two-signal mechanism using protein ligand and receptor pairs. Signal one results from recognition of antigenic peptide-MHC (major histocompatibility complex) complexes by the antigen-specific T-cell receptor. The second costimulatory signal is delivered by B7-1/B7-2 ligands on the antigen-presenting cell to the T cell through the CD28 receptor [1]. In contrast, engagement of CTLA4, a receptor that is structurally homologous to CD28, by B7-1 or B7-2 ligands results in the down-modulation of T-cell activation. Although CD28 and CTLA4 seem to mediate opposing function in T-cell physiology, these two receptors are related in structure and have been found to colocalize to one ~ 150-kb yeast artificial chromosome (YAC) clone [2].

A CD28-like receptor, ICOS [3], and its B7-like cognate ligand, GL50, were identified in both mouse and human systems ([4]; also known as B7RP [5] and B7h [6]). CD28 and ICOS exhibit protein sequence identity of ~ 24%, just as the GL50 proteins also share ~ 24% sequence identity with B7 proteins. Despite structural similarity, neither GL50 nor ICOS is likely to use the B7:CD28/CTLA4 costimulatory pathway because of the inability of GL50 to bind CD28/CTLA4 proteins and the inability of B7 proteins to bind ICOS-receptors [7]. *In vitro* analysis of ICOS mediated T-cell costimulation revealed that ICOS engagement resulted in enhanced T-cell proliferation and Th-2 cytokine production. Blockade of the ICOS pathway by addition of ICOS-Ig to MLR (mixed lymphocyte reaction) or tetanus toxoid recall response assays resulted in decreased T-cell proliferation [8]. Transgenic mice expressing ICOS–ligand exhibited an increase in B-cell germinal center size and enhancement of immunoglobin production [5], suggesting that overexpression of the ligand may influence B-cell development. These data are consistent with the model of the ICOS receptor serving as a pivotal signaling molecule involved with T-cell and B-cell proliferation and differentiation.

0888-7543/01 $35.00

doi:10.1006/geno.2001.6655, availa ine at http://www.idealibrary.com on IDEAL

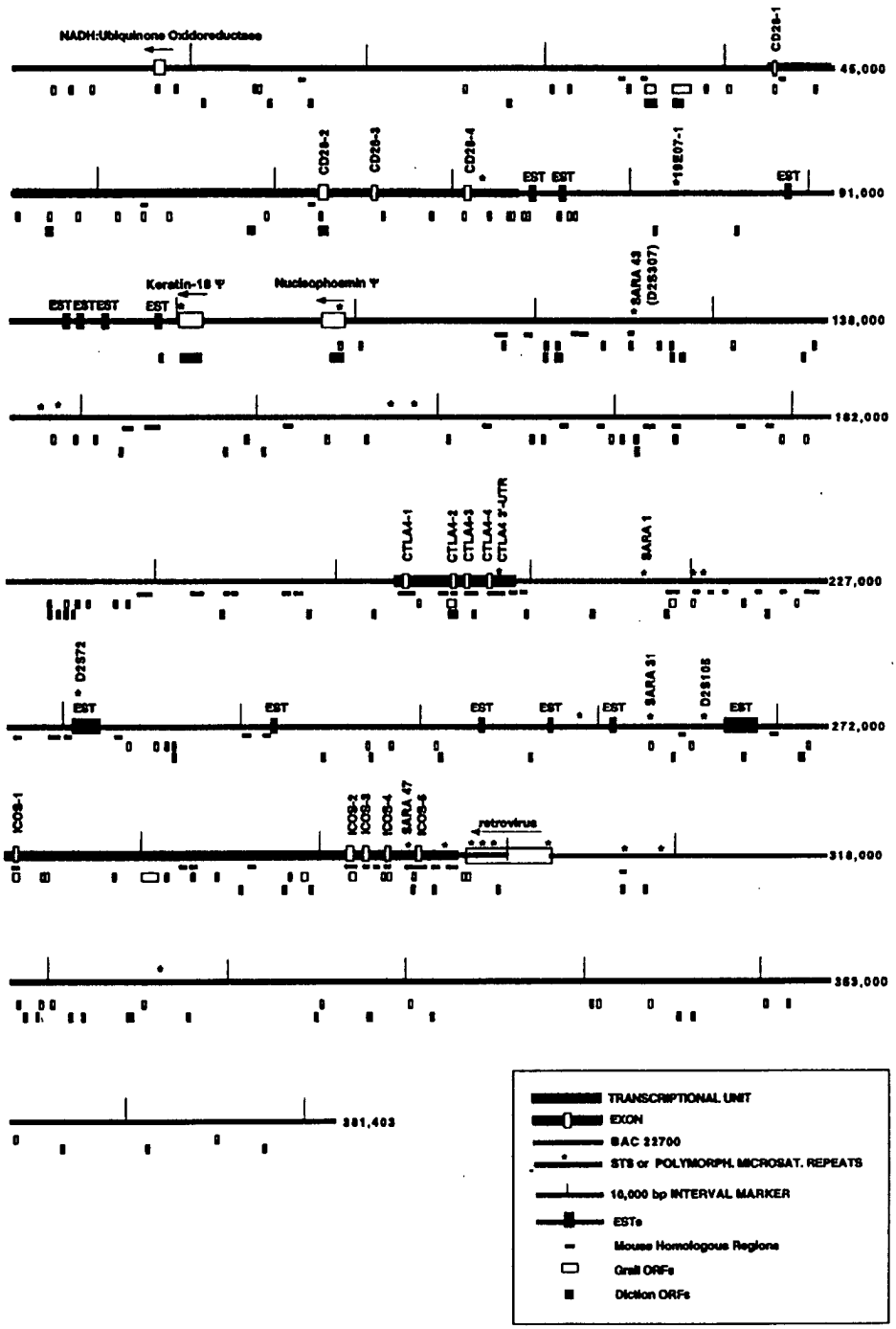

FIG. 1. Sequence diagram of the human 2q33 costimulatory receptor region. The position of the sequence line is indicated as nucleotides displayed. Stippled line represents human BAC clone 22700 sequence. Coding sequences of NADH:ubiquinone oxidoreductase, keratin-18 pseudogene, nucleophosmin pseudogene, EST-like sequences, retroviral elements, *CD28* (4 CDS), *CTLA4* (4 CDS), and the *ICOS* (5 CDS) receptors are displayed as open boxes on the sequence line. Black bars beneath the sequence line indicate regions of mouse sequence homology (> 35 bp, > 70% identity) based on limited sequencing of mouse BAC clone 23114 syntenic to human BAC clone 22700. White boxes below the sequence line indicate predicted ORFs by Grail; gray boxes indicate predicted ORFs by DiCTion. Sequences with homologies to GenBank STS and microsatellite repeats are marked as asterisks. Polymorphic microsatellite repeats used in this study are indicated as SARA 43, SARA 1, SARA 31, *CTLA4* 3′ UTR, and SARA 47.

The degree of structural similarity found between the ICOS receptor and CD28 combined with the genetic proximity of *CD28* to *CTLA4* led us to examine whether the ICOS receptor gene colocalized to the 2q33 region of the human genome. The phenomena of immunological gene clustering has been demonstrated with the initial mapping of 5q31 followed by noncontiguous sequencing of a 680-kb region revealing the presence of IL-3, IL-4, IL-5, IL-13, and GM-CSF genetic loci [9]. These genes do not share high similarity at the nucleotide or protein sequence level, but have been demonstrated to share some overlap in tertiary structure and biological activities. The sequence determination and assembly of chromosome 21 revealed an additional immunological gene cluster containing the type II cytokine receptor genes including *IFNAR2*, *IL10R2*, *IFNAR1*, and *IFNGR2* [10].

In a previous study, we demonstrated extensive sequence identity in 9 kb shared between the human and mouse *CTLA4* loci with similarity values of an average of 68% across exon boundaries, including intron sequences and 5′ and 3′ untranslated sequences [7]. These results were highly reminiscent of the conserved genomic patterning (70% average) detected between the mouse and human T-cell receptor gene, spanning a distance of 100 kb. We postulated that by comparative genomic analysis, receptors associated with immunological signaling and costimulation may exhibit a high degree of sequence homology that may extend beyond the *CTLA4* locus, thus predicting that other receptor genes in the proximity of *CTLA4* may exhibit mouse/human genomic sequence conservation. Here we confirm this hypothesis by sequencing human bacterial artificial chromosome (BAC) clones associated with 2q33 and its syntenic region in mouse and show that *ICOS* is localized near

CD28/CTLA4, forming an immunological costimulatory receptor cluster spanning 381 kb, with a high degree of organizational and functional similarity between human and mouse *ICOS*. We further demonstrate the induction of *CTLA4* and *ICOS* transcription by a novel microarray application in which genomic DNA is biologically interrogated for transcriptional activity.

RESULTS

Physical Mapping, Genomic Sequencing, and Assembly of 2q33 Costimulatory Receptor Cluster

To determine the degree of overlap and distance between *CTLA4*, *CD28*, and *ICOS*, we isolated six independent BAC clones by hybridization to costimulatory receptor cDNA probes. Of the six separate BAC clones, two exhibited hybridization with *CD28*, two with *CTLA4*, one with *ICOS*, and one with both *CTLA4* and *ICOS*. Each BAC clone was end-sequenced and PCR primer sets were designed to examine BAC clone overlap. Overlapping PCR sets were detected between BAC clones resulting in a hypothetical map of the costimulatory receptor region clustered in the order of *CD28*, *CTLA4*, and *ICOS*. Threefold shotgun sequencing of the clone 22700 library resulted in the generation of 1151 end-reads collapsing into 70 contigs spanning approximately 170 kb. Twofold sequencing of the clone 22606 and 22608 libraries generated 960 sequences collapsing into 107 contigs spanning 130 kb and 960 sequences collapsing into 111 contigs spanning 107 kb, respectively. Mouse BAC clone 23114 was sequenced twofold generating 767 end-read sequences collapsing into 143 contigs spanning 131 kb. Big-Dye primer sequencing was carried out directly on BAC clone DNA using primers designed from the sequences flanking gapped sites to close selected gaps in sequence.

BAC clones were end-sequenced and PCR primer sets designed specific to each BAC end. Amplification of each BAC clone with the complete set of PCR primers resulted in amplification patterns corresponding to the genomic organization of the costimulatory receptors. Starting and ending positions based on subsequent sequence data are indicated for each BAC clone (N.D., not determined): BAC 22606 (N.D. to 66,887), BAC 22607 (N.D. to 167,094), BAC 22701 (74,706 to 278,563), BAC 22699 (84,599 to 239,485), BAC 22700 (119,296 to 300,949), BAC 22608 (233,866 to 381,403).

When necessary, we used overlaps to publicly available genomic data to position contigs, especially PAC clone p61e2 (acc. no. AF225900), bridging the 52,408-bp gap between nt 66,888 and nt 119,295. Merging BAC clones with existing sequences resulted in one contiguous sequence of 381,403 bp initiating 42,570 bp upstream of *CD28* and ending 85,985 bp downstream of *ICOS* (Fig. 1).

Genomic Organization of 2q33 Genes, Homologues, STSs, and ESTs

We identified 20 potential protein coding elements within the 381-kb costimulatory receptor region with sequences exhibiting either identity to or homology with known genes or ESTs (Table 1 and Table 2): NADH:ubiquinone oxidoreductase homologue, CD28 (NM_006139), keratin-18 pseudogene, nucleophosmin pseudogene, CTLA4 (NM_005214), Unigene HS.30542 homologue, ESTs, ICOS (Genseq V53199), and an element similar to many human endogenous retrovirus type H with associated 5′ and 3′ LTRs (RTLV-H2, M18048; among others). Based on a recent mapping study of 2q31–q33, the three receptor loci within this region are situated on the chromosome with *CD28* being the most centromeric and markers, now known to be near *ICOS*, being the most telomeric [11]. In addition, 22 STSs were identified upon BLAST search of this compiled region of 2q33, of which 4 correlated to endogenous retroviral sequence (Table 3). The commonly used genetic markers for 2q33, *D2S307* (SARA 43), *D2S72*, *D2S105*, and 19E07-1, were contained within the sequence presented here. Because HERV-H elements are found in ~ 1000 copies in the genome, it remains to be determined if these four STSs are specific for the element described here. Based on human *ICOS* cDNA sequence data, the organization of the *ICOS* locus was determined to be five coding sequences spanning 22,758 bp from the initiation codon of exon 1 to the termination codon of exon 5, unlike the four-exon structure of both *CTLA4* and *CD28*. *ICOS* exon 5 encoded the smallest coding sequence, represented by only four amino acids [(D)-V-T-L] followed by a stop codon. In other respects, exons 1–4 parallel the genomic organization of *CTLA4* and *CD28* with exon 1 encoding the leader sequence, exon 2 encoding the extracellular Ig-V-like domain, exon 3 encoding the transmembrane domain, and exons 4 and 5 encoding the the cytoplas-

TABLE 1: Summary of 2q33 sequence information

Feature type	Number	Total length	Ave length	SD	Proportion of analyzed region
Simple repeats	353	9604	27	27	2.52%
Complex repeats	368	60536	151	68	15.87%
Grail ORFs	118	18799	159	130	4.93%
DiCTion ORFs	70	17476	250	110	4.58%
Syntenic mouse >35 bp	70	8497	121	124	-
Costimulatory receptors (transcribed unit)	3	62285	-	-	16.33%
Other genes/pseudogenes/EST	17	15382	-	-	4.03%
Sequence tagged sites	22	9241			2.42%

TABLE 2: Feature table of the human costimulatory receptor region of chromosome 2q33

Receptor	Position Start	Position End	Size	Intron	Size	Gene/EST	Position Start	Position End	Size	Reference	Notes
CD28 5'UTR	42348	42569	222			NADH :ubiquinone oxidoreductase homolog	7838	8329	491	AF201077	
CD28 CDS-1	42570	42621	52	CD28 intron 1	19883	EST	74209	74682	473	AA311148	from Jurkat T-cell library
CD28 CDS-2	62505	62861	357	CD28 intron 2	2678	EST	75932	76379	447	N20227	from Melanocyte library
CD28 CDS-3	65540	65664	125	CD28 intron 3	5010	EST	88605	88873	268	AA663852	from schizo. brain library
CD28 CDS-4	70675	70803	129			EST	93458	93983	525	AA744591	vicinity of multiple repeat elements.
CD28 3' UTR	70804	73724	2921			EST	94424	94744	320	H89084	multiple EST hits
						EST	95762	96257	495	AW237774	multiple EST hits
CTLA4 5' UTR	203644	203799	156			EST	98855	99173	318	L44301	human thymus library
CTLA4 CDS-1	203800	203908	109	CTLA4 intron 1	2534	Keratin 18 pseudogene	100130	101424	1294	M26325, #NM_000224	multiple stops
CTLA4 CDS-2	206443	206790	348	CTLA4 intron 2	444	Nucleophosmin pseudogene	108193	109455	1262	M26697, #NM_006993	multiple stops
CTLA4 CDS-3	207235	207346	112	CTLA4 intron 3	1218	EST homolog	230519	232134	1615	R91770, AW474005, AI434725	vicinity of multiple repeat elements
CTLA4 CDS-4	208565	208669	105			EST homolog	241762	242097	335	AW238656, AL037926, AI905493	possible distant L1 repeat
CTLA4 3'UTR	208670	209793	1124			EST homolog	253467	253534	67	N73819, AI801031, AW079941	vicinity of multiple repeat elements
						EST homolog	257288	257506	218	Unigene cluster homolog HS.30542	cDNA clusters from multiple tissue sources
ICOS 5' UTR	272636	272660	25			EST	260890	261082	192	AA663871	schizo. brain library
ICOS CDS-1	272661	272718	58	ICOS intron 1	18753	EST homolog	267282	269005	1723	AA558770, AA054182, T90825	vicinity of multiple repeat elements
ICOS CDS-2	291472	291807	336	ICOS intron 2	685	Endogenous retrovirus	297760	303099	5339	AF139170, PIR:A44282	79% identity to some retroviral elements
ICOS CDS-3	292493	292599	107	ICOS intron 3	1032						
ICOS CDS-4	293632	293716	85	ICOS	1689						
ICOS CDS-5	295406	295419	14	intron 4							
COS 3' UTR	295420	297393	1974								

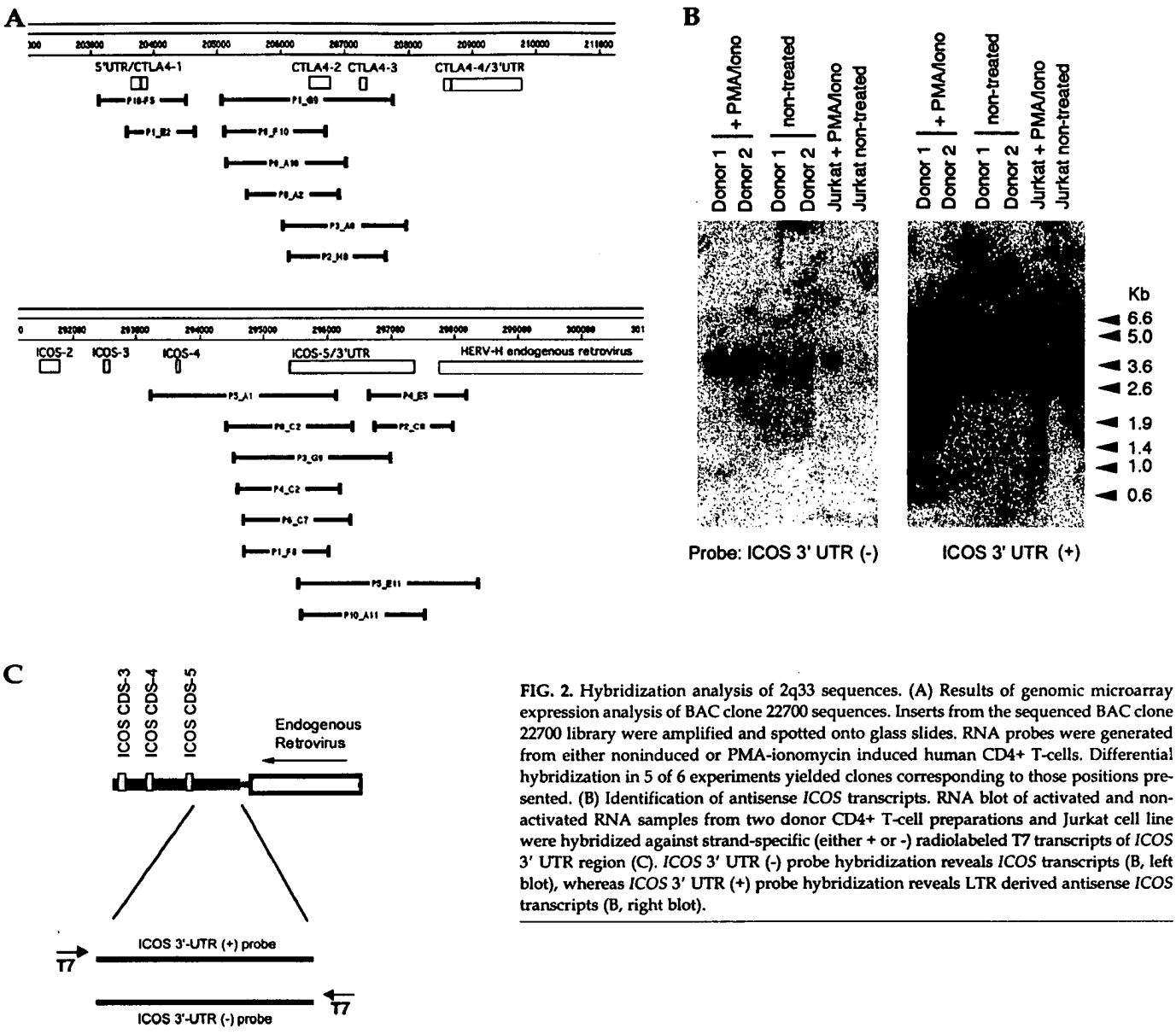

FIG. 2. Hybridization analysis of 2q33 sequences. (A) Results of genomic microarray expression analysis of BAC clone 22700 sequences. Inserts from the sequenced BAC clone 22700 library were amplified and spotted onto glass slides. RNA probes were generated from either noninduced or PMA-ionomycin induced human CD4+ T-cells. Differential hybridization in 5 of 6 experiments yielded clones corresponding to those positions presented. (B) Identification of antisense *ICOS* transcripts. RNA blot of activated and non-activated RNA samples from two donor CD4+ T-cell preparations and Jurkat cell line were hybridized against strand-specific (either + or -) radiolabeled T7 transcripts of *ICOS* 3′ UTR region (C). *ICOS* 3′ UTR (-) probe hybridization reveals *ICOS* transcripts (B, left blot), whereas *ICOS* 3′ UTR (+) probe hybridization reveals LTR derived antisense *ICOS* transcripts (B, right blot).

mic domain. All three costimulatory receptors shared similar pattern of intron size distribution in which intron 1 > intron 3 > intron 2. *ICOS* appeared to be more similar in genomic organization to *CD28*, with *ICOS* intron 1 spanning 18.7 kb compared with *CD28* intron 1 spanning 19.9 kb, versus *CTLA4* intron 1 spanning 2.5 kb.

Computer Assisted Prediction of Open Reading Frames

The 381-kb costimulatory receptor locus was analyzed by the open reading frame (ORF) prediction programs DiCTion and GRAIL to assess the potential of other sequences in this region to encode gene products (Fig. 1 and Table 1). DiCTion analysis of the costimulatory receptor region resulted in the prediction of 70 ORFs with a cumulative length of 17,476 bp, of which 5 ORFs represented repetitive *Alu* sequences. Coding sequences representing *CD28* exon 2 and *CTLA4* exon 2, keratin-18, and nucleophosmin pseudogenes were predicted by DiCTion. DiCTion did not predict sequences encoding ICOS. Of the remaining ORFs, two were localized to intron 1 of *CD28*, and single ORFs were predicted in intron 3 of both *CTLA4* and *ICOS* receptor loci. Assuming that the predicted intronic ORFs are false positives, these results suggest that up to 56 potential DiCTion ORFs remain in this region of 381 kb. GRAIL analysis generated more potential ORFs than DiCTion, with a total of 118 segments and a cumulative length of 18,799 bp (Table 1). GRAIL predicted some ORFs containing *CD28* (CDS-1, CDS-2, CDS-4), *CTLA4* (CDS-2), and *ICOS* (CDS-1, CDS-2, CDS-4), but neither GRAIL nor DiCTion was successful in predicting the complete set of exonic sequences from any receptor. Moreover, both programs predicted ORFs

doi:10.1006/geno.2001.6655, available online at http://www.idealibrary.com on IDEAL

TABLE 3: BLAST sequence homologies of 2q33 to STS

	STS Accession no.	Start	Stop
STS	L18270	31270	31614
STS	G06727	71001	71556
STS	G09915	82496	82773
STS	G33149	100138	100487
STS	L17895	108738	109273
STS	G54502	125332	125844
STS	G54505	137509	137959
STS	G31888	138782	138954
STS	G54503	157440	157808
STS	G60294	158705	159022
STS	G31884	219881	220435
STS	G52661	220517	220915
STS	G54504	258807	259508
STS	M98994	265922	266199
STS	G54507	296584	297060
STS*	G49372	298172	298384
STS*	G14647	298907	299257
STS*	G41957	299354	299985
STS*	AF191977	302665	303099
STS	G16447	307078	307448
STS	G54508	308944	309478
STS	G04861	326312	326460

*Denotes BLAST hits to retroviral sequences.

in known intronic sequences. For example, in *CD28* intron 1, GRAIL predicted eight ORFs, whereas DiCTion predicted one ORF. Although it has been reported that CD28 may be expressed as alternatively spliced products [12], it has not been demonstrated that intronic sequences described here contribute to the final products of known isoform variants. When DiCTion and GRAIL outputs were compared; 13 predicted ORFs were found in common to both. Of these, three correspond to the known sequences *CD28* CDS-2, *CTLA4* CDS-2, and EST M26697.

Genomic Microarray Expression Analysis (GMEA)

To examine whether differentially transcribed genes within this genomic region could be detected, we interrogated the sequenced BAC 22700 subclone library collection by genomic microarray expression analysis. The previously sequenced plasmid library DNA samples were amplified by PCR, the amplified DNA products were spotted onto glass slides, and hybridization was carried out with total RNA from either nonstimulated or PMA-ionomycin treated CD4+ T cells. Of the starting 864 plasmid subclones, 620 amplified products were recovered and analyzed, resulting in 18 clones showing differential hybridization in five of six replicate experiments (three slides each with duplicate spots). Eight clones corresponded to sequences within the *CTLA4* locus, seven clones corresponded only to the *ICOS* 3′ UTR, and three clones corresponded to both *ICOS* 3′ UTR and endogenous retroviral sequences immediately 3′ of *ICOS* (Fig. 2A). It must be noted that hybridization of cDNA against genomic DNA would preferentially occur between target sequences of longer length (exon 2 and 3′ UTR of *CTLA4* and *ICOS*); thus the degree of hybridization to microarrayed spots containing only short CDS flanked by non-differentially expressing intronic sequences could be lower. Indeed, the differential hybridization detected to *ICOS* was to the region corresponding to the longest transcribed unit, the 2-kb 3′ UTR. No clones other than *CTLA4*, *ICOS*, and retrovirus immediately downstream of *ICOS* were found to be induced, suggesting that the stringency of the experimental conditions used in this study was sufficient for detecting transcriptionally induced genes while effectively eliminating nonspecific background hybridization generated by genomic and plasmid DNA.

To determine whether hybridization to *ICOS* and retroviral sequences reflected transcription from the *ICOS* promoter or whether this differential signal reflected transcripts from the endogenous retrovirus proximal to the *ICOS* locus, we carried out RNA blots to determine transcript orientation from this region. To rule out cross-hybridization to repetitive sequences, a BLAST search was performed using *ICOS* 3′ UTR sequences adjacent to the endogenous retrovirus. No repetitive DNA was detected, hence, this sequence was subcloned in both orientations into separate T7-promoter-bearing vectors to generate strand-specific radiolabeled probes. RNAs from two donor CD4+ T cells and Jurkat T-cell line preparations, cultured in either the presence or the absence of PMA-ionomycin activation, were fractionated, blotted, and hybridized to the *ICOS* 3′ UTR sense or antisense probe (Fig. 2B). With the *ICOS* antisense probe, a clear hybridization signal was observed for activated samples but not for non-activated samples. Hybridization with *ICOS* sense probe also revealed two regions of clear hybridization signals in all samples examined: one discrete band at approximately 6.5 kb and one non-discrete band at ~ 3–4 kb. These results suggest that the retroviral LTR promoters 3′ of *ICOS* are transcriptionally active and are responsive to cell activation. The 6-kb band seemed to be preferentially induced on activated CD4+ T-cells while being constitutively expressed in both Jurkat cells samples. The 3- to 4-kb band appeared to be expressed in all samples examined regardless of activation state. Because these retroviral transcripts may be derived from either the 5′ LTR or the 3′ LTR viral promoter, at least two potential sets of transcripts may be detected. With the presence of eight canonical polyadenylation signals (AATAAA) within the 7.5 kb upstream from the *ICOS* 3′ UTR, it is not possible to correlate promoter activity with observed transcript size at this time.

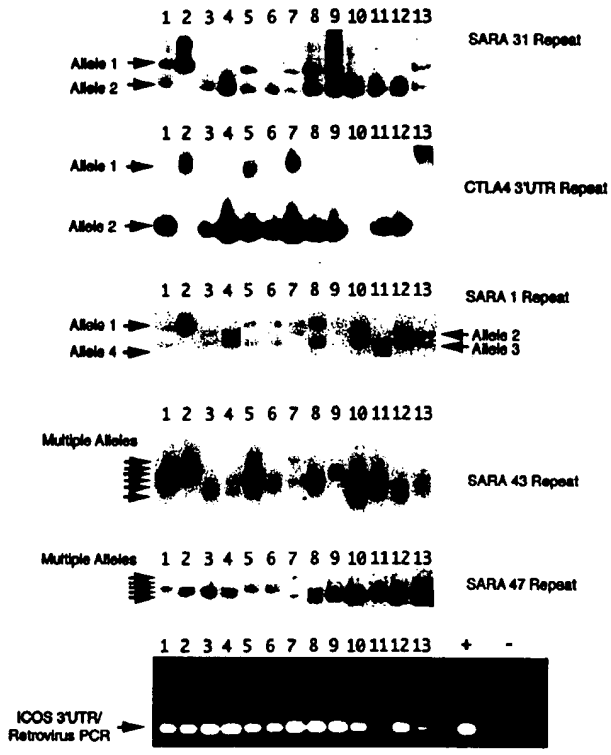

FIG. 3. Identification of polymorphic microsatellite repeats within BAC clone 22700. Amplification of repeats SARA 31, *CTLA4* 3′ UTR, SARA 1, SARA 43, and SARA 47 followed by denaturing PAGE electrophoresis and autoradiography revealed polymorphic PCR products. Two alleles were detected in SARA 31 and *CTLA4* 3′ UTR; four alleles were detected in SARA 1, and more than five alleles were detected in both SARA 43 and SARA 47 amplification reactions.

Analysis of Microsatellite Polymorphisms

Polymorphisms in the 3′ UTR of *CTLA4* have been linked to several autoimmune genetic diseases. To identify additional markers in this region that may also serve to refine the associations between genetic diseases and the costimulatory receptor region of 2q33, we analyzed 25 microsatellite repeat sequences in the BAC 22700 clone for the presence of repeat unit polymorphisms. Genomic DNA PCR amplification of 13 individuals revealed four microsatellites, corresponding to di-, tri-, and hexanucleotide repeats, that demonstrated allelic polymorphisms upon analysis by denaturing acrylamide gel electrophoresis (Fig. 3). Of the four polymorphic microsatellite repeats examined, repeat SARA 31 (nt 263,177–263,211; [ATTTTTT]n6) was represented by two alleles, repeat SARA 1 (nt 217,444–217,492; [TATC]n12) was represented by four alleles, and SARA 43 (nt 125,845–125,892 [GT]n24, homologous to sequences within *D2S307*) and SARA 47 (nt 295,275–295,326; [GT]n15) seemed to be highly polymorphic with at least six different alleles within 13 individuals examined. Analysis of the 13 individuals for the polymorphisms associated with the known *CTLA4* 3′ UTR (nt 209,177–209,216; [AT]n40) microsatellite repeat demonstrated two alleles. Compilation and comparison of the four polymorphic microsatellite alleles found in these individuals revealed no shared allelic combination, indicating that this set of four polymorphic markers may be effectively applied to the high-resolution discrimination of genetic associations of disease states linked to the costimulatory receptor region. For a positive amplification control, a primer set was used corresponding to nt 297,362 to 297,388 (forward primer) and 297,934 to 297,907 (reverse primer) corresponding to the 3′ UTR of *ICOS* and to the 3′ LTR of the HERV-H. Amplification of the 13 individuals with this set of primers resulted in a single predicted band at ~ 400 bp indicating the presence of this segment of DNA across the panel examined.

Cross-species Comparison of *ICOS*

The generation of the complete sequence for the human *ICOS* locus along with the partial sequencing of the mouse *Icos* locus allowed the cross-species comparison of genomic coding and noncoding sequences in this region (Figs. 4A, 4B, and 4C). Limited gap closure of the mouse *Icos* locus by primer walking resulted in the assembly of one contiguous sequence spanning CDS-2 to CDS-5 and flanked by 2265 bp of intron 1 and 1415 bp of 3′ untranslated/genomic DNA. Dot-plot comparison analysis of the human genomic region was performed with the syntenic genomic region from mouse starting from 2265 bp upstream of mouse CDS-2 to 1414 bp downstream from mouse CDS-5 (Fig. 4B). Allowing for gaps, diagonals representing a minimum of 60% sequence identity were clearly observed in this aligned region; most notably, a diagonal was detected extending 3′ of CDS-5 for 2.4 kb. A similarity plot of the gap-corrected sequence alignment of this region resulted in approximately 60% sequence identity over 6.4 kb of aligned sequence. The highest peaks of sequence similarity (~ 80% identity) were clearly detected for CDS-2, CDS-3, CDS-4, and CDS-5. Intron 2 and intron 3 had a lower similarity score (~ 45%) owing to the presence of gaps formed by the alignment process. Gaps in alignment represented by valleys (< 30% identity) were generally composed of repetitive sequences presented in only one species. We found seven peaks of high sequence identity (> 70%) in noncoding regions of intron 4 and the 3′ UTR region starting from 1 kb upstream to 2.4 kb downstream of CDS-5. The sequence conservation in the *ICOS* intron 4 was especially striking, as evidenced by the presence of the SARA 47 microsatellite in both mouse and human sequences. The SARA 47 (GT)n24 intron 4 microsatellite repeat was located 88 bp 5′ of human *ICOS* exon 5, whereas a similar (GT)n48 intron 4 microsatellite repeat was discovered 66 bp 5′ of mouse *Icos* exon 5.

Sequences flanking *ICOS* CDS-1 revealed two zones of high similarity between mouse and human genomic DNA

doi:10.1006/geno.2001.6655, available online at http://www.idealibrary.com on IDEAL

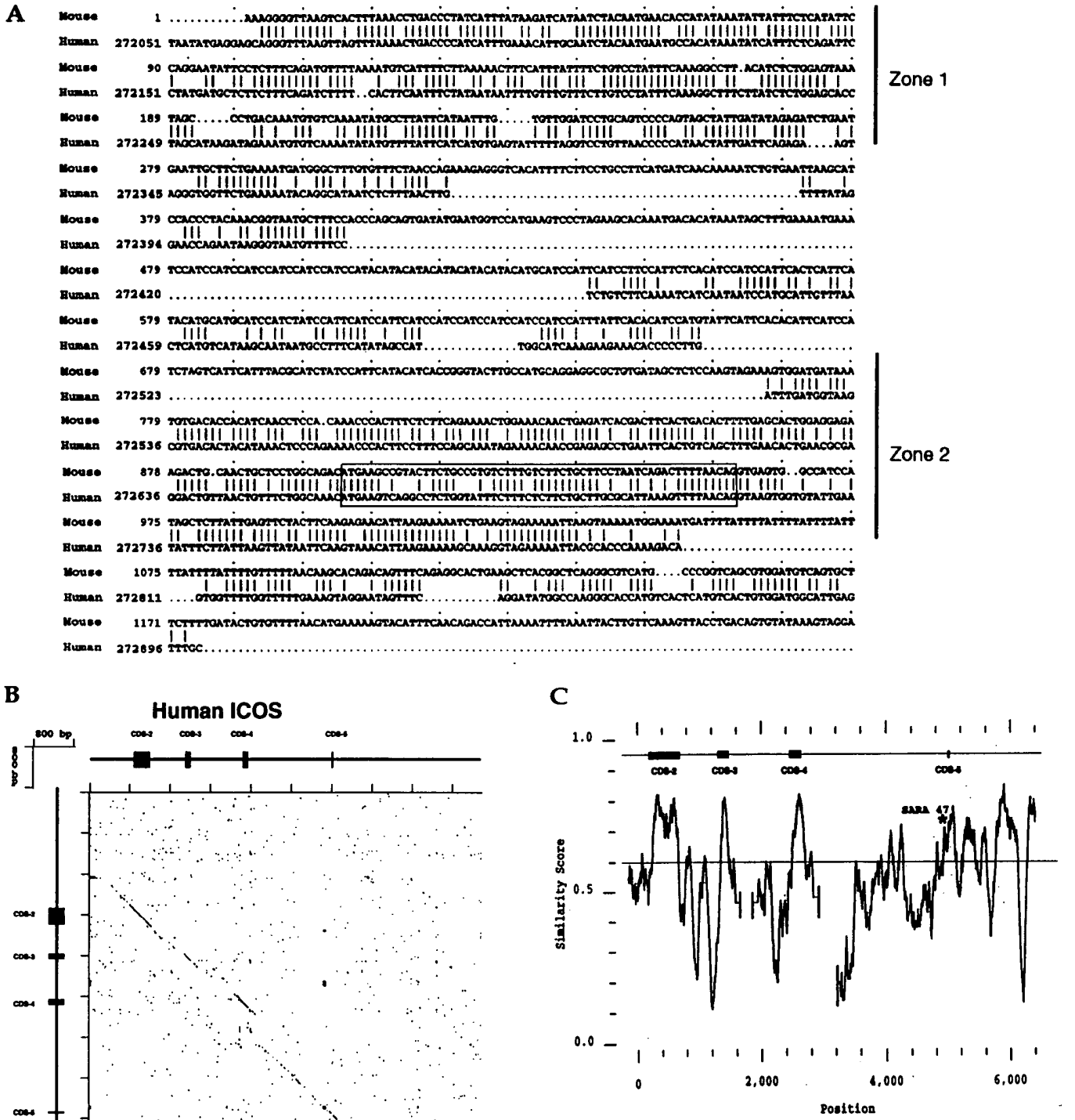

FIG. 4. Sequence alignment between mouse and human *ICOS* genomic DNA. (A) GAP alignment of regions flanking CDS-1 (boxed) revealed two zones of sequence homology (as shown) separated by a ~ 250-bp mouse-specific repetitive DNA region. (B) Dot-plot alignment of human and mouse *ICOS* genomic regions including CDS-2 to CDS-5. Homologies greater than 60% identity over a 20-bp window are displayed. (C) Similarity plot of consensus sequence derived from GAP alignment between human and mouse *ICOS* genomic regions displayed in (B). Breaks in similarity index indicates presence of nonconserved repetitive sequences. Aligned consensus coding sequences are indicated in the top line, while the location of the conserved SARA 47 microsatellite repeat is denoted by an asterisk.

(Fig. 4A). The first zone of high sequence identity was a 317-bp region with 72% sequence identity to mouse sequences located 276-bp upstream from initiation methionine at nt 272,661. The second zone was a 269 bp region with 75% sequence identity immediately flanking and including CDS-1, starting from 134 bp upstream of the initiation methionine to 75 bp downstream from the start of intron 1. The intervening gap (human = 143 bp, mouse = 448 bp) between zone 1 and zone 2 was due to a G-deficient tract of DNA unique to mouse sequence and populated with numerous low-complexity TCCA, TACA, and TTCA repeats. Assuming that transcriptional control regions are conserved between mouse and human, it is likely that sequences in either zone 1 or zone 2 are responsible for transcriptional control of *ICOS* expression. The full-length human *ICOS* cDNA (Genseq no. V53199) reveals 25 bp of 5′ UTR before the initiation codon, but whether this cDNA clone represents the actual transcription start site remains to be determined. Neither mouse nor human *ICOS* zone 2 contains the conventional TATA promoter motif, suggesting that the transcriptional start site is likely to be in zone 1, which contains multiple TATA sites. Analysis for conserved transcription factor binding sites located in both zone 1 and zone 2 by Transfac database search (http://www.cbil.upenn.edu/tess/index.html) revealed no T-cell-specific control elements shared between mouse and human sequences. A single potential NFAT-1 site was found in mouse zone 1 along with numerous non-T-cell-specific sites (for example, AP-1, AP-2, Pu.1, GATA-1, c-Jun, Gal4, and others).

The extent of sequence conservation within the intergenic region encompassing *CTLA4* and *ICOS* was examined by a comparative genomic survey of a twofold sequenced syntenic mouse BAC clone comprising 143 non-contiguous sequences aligned with the repeat-masked (DUST) human 381-kb sequence using SIM4. For regions greater than 34 bp in length, 71 alignments were found with identity scores averaging 81%. When human sequences between nt 100,000 and 301,000 were examined, repetitive sequences comprised 36,621 bp, leaving a total of 164,379 bp of potential structural or transcribed DNA. Within this region, SIM4 mouse homologies totaled 8531 bp theoretically corresponding to roughly 5% of the *CTLA4/ICOS* region. Given the limited degree of mouse BAC clone sequence coverage, only 131 kb of data were generated with the potential for an additional missing 28 kb in "unfilled" gaps, leaving the sequence determination of the syntenic mouse region approximately 80% complete. Based on the 5% homology estimated between mouse genomic DNA syntenic and shared with human BAC clone 22700, it is not likely that extensive sequence similarities span the intergenic region between *CTLA4* and *ICOS*, but rather similarities are composed of smaller stretches of homologous DNA within this region. It remains to be determined whether these stretches of homologous genomic DNA are involved with transcriptional control or whether they encode other peptide domains common to both species.

Discussion

Molecular Phylogeny of Costimulatory Receptors

We have shown that *CD28*, *CTLA4*, and *ICOS* are closely linked, with 252 kb separating the initiation codon of *CD28* and the termination codon of *ICOS*, and that the mouse syntenic genomic region shares similar organization with areas of high sequence similarity. In contrast, the CTLA4-like inhibitory receptor PD-1, mapped at 2q37.3 [13], was not present in this assembly. Within the assembled human 2q33 costimulatory receptor region, 2.5% (9604 bp) was composed of low-complexity repeats, whereas 15% was composed of high-complexity repeats, leaving ~ 82% DNA available to potentially encode transcriptionally active DNA. In addition to *CD28*, *CTLA4*, and *ICOS*, the 381-kb costimulatory receptor region contained sequences with homology to entries found in GenBank. Two common pseudogenes were found corresponding to keratin-18, an intermediate filament protein associated with carcinomas [14], and nucleophosmin, a ribosomal assembly protein [15]. Although we have not assessed whether these pseudogenes are transcribed or whether they may have biological significance, the potential of these pseudogenes to produce functional proteins is low. Both pseudogenes contain initiation codons but have interspersed multiple nonintronic termination codons in all three reading frames. One potentially novel gene was identified with high sequence similarity to NADH:ubiquinone/oxidoreductase MLRQ, an 81-amino-acid protein that functions in complex I of the mitochondrial respiratory chain, catalyzing the following reaction: NADH + ubiquinone = NAD + ubiquinol. The MLRQ subunit, defined by the first four residues of the protein, is found in several species, presumably sharing a common role in mitochondrial respiration [16]. It is not known whether the novel NADH:ubiquinone oxidoreductase presented here (79% identity to MLRQ form) is transcribed and represents a novel member of this protein family. It is interesting to note that all three costimulatory receptors share similar transcriptional orientation. In contrast the pseudogenes, NADH:ubiquinone oxidoreductase, and HERV-H element are all positioned in the opposite orientation. The shared directionality of the costimulatory receptors is analogous to the unidirectional pattern revealed for the type II cytokine receptor cluster in the sequenced chromosome 21 [10]. It has been hypothesized that a common orientation between structurally related proteins within the genome may reflect the outcome of initial gene duplication events.

Comparative genomic analysis of the *ICOS* locus revealed certain peculiarities in sequence divergence. The presence of the highly polymorphic SARA 47 microsatellite in both mouse and human sequences (Fig. 4C) suggests that this repetitive sequence is both ancient and well conserved, as little degeneration of repeat unit periodicity was observed. This is surprising, considering the presence of a unique tract of tetranucleotide repeats found in the 5′ UTR, 200 bp upstream of the mouse *Icos* initiation methionine that is of low complexity

and cryptic organization. The presence of these cryptic 5′ repeats unique to the mouse genome suggests that they are of a more recent origin than the shared SARA 47. Paradoxically, the degree of degeneracy within the 5′ repeat is higher than SARA 47, suggesting that these sites are either ancient or have not been preserved from mutation. However, it must be noted that the origin of simple repetitive sequences is unclear and it is questionable if either of these repetitive regions confers any function in cellular physiology. We have previously reported the similar preservation of a shared microsatellite repeat (hR1) between mouse and human *CTLA4* intron 3 DNA [7]. Analysis of the human *CTLA4* hR1 microsatellite repeat within the limited DNA samples presented in this study, however, has not yet yielded polymorphic alleles (data not shown). We have previously suggested that preservation of noncoding sequences within *CTLA4* between species may indicate the presence of a mutational constricting mechanism that limits the accumulation of nucleotide variation across all costimulatory receptors. Indeed, although the ICOS receptor genomic sequences shares a high degree of sequence similarity (~ 60%) between mouse and human in certain sections, it is less than the similarity observed between mouse and human *CTLA4* genomic sequences (> 70%), in which extensive sequence similarity is found across exon boundaries with sections exhibiting similarities greater than 90% identity. In this study, analysis of syntenic mouse contigs analyzed for homology to human sequences revealed that conserved contigs (> 70% homology) only constituted ~ 5% of the total contigs assembled, suggesting that sequence conservation in intergenic regions around the *CTLA4* and ICOS loci may not be extensive. We found that *CD28*, *CTLA4*, and *ICOS* CDS sequences exhibit cross-species sequence identity at 67%, 74%, and 69%, respectively. This comparative analysis evokes several possibilities that may account for the preferential maintenance of *CTLA4* genetic structure over that of *ICOS*: stronger mutation-constraining mechanisms may exist closer to the *CTLA4* locus than the *ICOS* locus; *ICOS* may have emerged before *CTLA4*; or other species-dependent mechanisms may exist to generate and maintain required costimulatory receptor sequence diversity. In a survey of human/rodent receptor/ligand pairs, it was found that host defense-related receptors and ligands exhibit the greatest degree of sequence divergence, with scores of ~ 65% amino acid identity [17]. In contrast the mean identity score of all other proteins analyzed in the available database ranged from 88% to 99%. It was hypothesized that the higher degree of sequence divergence seen for immunological proteins was due to molecular mimicry of pathogens masking as cytokines or their receptors, thus providing a basis for driving selection of increasingly divergent immunological ligand/receptor pairs. From studies of mouse/human cytokines (data not shown), we have found that receptor identities of greater than 60% between species generally correlate with cross-species-reactive cytokine activity. As expected, the conservation of 69% protein sequence identity deduced from the ICOS CDS found between the two species was adequate for cross-species ligand-receptor binding reactivity in flow cytometric based assays using human and mouse ICOS–Ig fusion protein reagents and ICOS–ligand cell transfectants (data not shown).

In a mouse/human comparative genomic analysis of a 1000-kb noncontiguous region containing the cytokine cluster 5q31, 90 conserved noncoding sequences greater than 100 bp in length and with greater than 70% identity were identified [9,18]. Of these, 15 were conserved across divergent mammalian species. The largest element, of 401 bp, was found to confer *cis*-acting transcriptional control activity, based on the transcriptional activity of human YAC deletion mutants in transgenic mouse cells. Deletion of the 401-bp element resulted in the decrease of human IL-4, IL-5, and IL-13 production upon mouse T-cell stimulation compared with nondeleted transgenic controls. These data suggest that cross-species conserved *cis*-acting control elements may have an impact on transcriptional events spanning distances upwards to 120 kb. *CD28*, *CTLA4*, and *ICOS* mRNAs are induced upon T-cell activation, analogous to the induction of the 5q31 cytokine cluster, leading to the possibility that transcriptional control elements may also occur in intergenic regions of 2q33. Based on the control element criteria described for 5q31, 25 conserved elements ranging from 100 bp to 605 bp with sequence identities of greater than 70% were identified in this chromosomal region. Given that most of these conserved genetic elements are not predicted to code for known proteins, it remains a possibility that these sequences may act as *cis*-acting transcriptional control elements.

Functional Genomics

With the advent of human genomic sequence data availability and the inability of computational methods to accurately predict all ORFs of encoded proteins, the necessity for novel experimental approaches to determine transcribed genomic DNA within biological settings will be essential. The GMEA process presented here is an experimental approach by which genomic regions of BAC clones can be readily fractionated and analyzed by fluorescence based hybridization to detect regions of differentially transcribed genomic DNA. With current microarray technology, cDNAs are commonly used as target hybridization substrates, with limitations in that cDNAs under-represented in the spotted libraries may not be detected. In another similar application, oligonucleotide arrays are made based on computer-predicted exons of genomic DNA [19]. In contrast to these methods, GMEA hybridization substrates are normalized targets, as they are derived from sheared, size-fractionated genomic DNA without prior sequence selection or bias. Further, GMEA addresses positional information of transcriptionally active positive clones, thus one could experimentally identify genomic regions bordering transcription initiation, intron/exon boundaries, and regions downstream of transcription termination. As presented here, this method provides positional information for transcriptional response elements located near the gene. GMEA may prove to be helpful in the future in uncovering novel genes and transcriptional

control elements to which genetic associations are mapped, but as yet have no corresponding protein product to examine in a physiological setting.

In this study, the spotting of 640 clones on a single slide resulted in approximately sixfold coverage of a 180 K BAC clone. However, with increased density of robotic spot placement, it is now possible to redundantly array the content of multiple BAC clones onto single glass slides. This would result in GMEA of more than 10 Mb of redundant genomic DNA using a single microarray (where 2 kb × 20,000 spots / 3× coverage = 13 Mb). In principle, using current technology, the human genome can be redundantly arrayed onto ~ 250 slides—approaching a number within the range to perform whole-genome GMEA. In this study, BAC clone 22700 microarrays revealed reproducible induction of *CTLA4* and *ICOS* transcripts and of retroviral transcripts immediately 3′ to *ICOS* in PMA-ionomycin activated T cells, consistent with reports that both *CTLA4* and *ICOS* are activation induced [1]. Because of the absence of other differentially hybridizing genomic fragments, these experimental results suggest that no other PMA-ionomycin activated T-cell genes are detected within the 181,654-bp section between positions 119,296 and 300,949 of the costimulatory receptor region. The stringency of the GMEA procedure was sufficient to avoid false identification of ~ 36 kb of simple and complex repetitive DNA in this BAC clone in addition to contaminating plasmid and bacterial sequences. Although hybridization to other genomic fragments was detected, the signal strength of these events was equivalent in both experimental and control samples, thus screening these data from differential expression analysis. Whether these non-differentially hybridizing sequences are transcribed products common to both activated and non-activated samples, or whether these represent background hybridization events remain to be determined in future experiments. One could envisage using GMEA to scan defined BAC subsequences potentially representing whole genomes to examine the correlation between expression patterns of transcribed DNA and loci attributed to genetic diseases. Further, RNA isolated from diseased and control states may be used to determine whether altered transcription levels of BAC-encoded gene products exist between pathological and normal samples.

Implications for Human Genetic Disease Linkage

Polymorphisms in 2q33 have been linked to numerous autoimmune genetic dieseases and familial primary pulmonary hypertension (PPH). Autoimmune diseases such as insulin-dependent diabetes, Grave's disease, Hashimoto's disease, and myasthenia gravis with thymoma [7] have been linked in certain populations to *CTLA4*. However, so far there is no evidence that any *CTLA4* polymorphism results in altered function of CTLA4, much less a mechanism leading to any of the pathologies associated with these disease states. The close proximity of the *ICOS* locus to *CTLA4* raises the possibility that genetic linkage of human autoimmune diseases to 2q33 could be due to variants in the *ICOS* locus, rather than the *CD28* or *CTLA4* loci. Chronic stimulation of ICOS receptor in ICOS-ligand transgenic mice revealed a phenotype consistent with B-cell hyperstimulation [5]. Grave's disease and type 1 diabetes share an etiology in which autoimmune antibodies are generated, suggesting a dysregulation of B-cell differentiation and antibody production. Although PPH is a clinical entity with some association with autoimmunity, familial PPH seems to be closely linked to mutations in the bone morphogenetic receptor-II gene, possibly located upstream of *CD28* [20].

A genetic study using STSs located within 200 kb of *CTLA4* revealed a high level of linkage disequilibrium with the onset of type 1 diabetes as reflected by the genetic locus IDDM12 [21]. Because of the limited coverage obtained from the number of markers used, the authors concluded that IDDM12 exists within roughly 100 kb of the *CTLA4* locus, but is not *CD28*. However, the IDDM12 physical map used in that study does not agree with the sequence determination presented here; the IDDM12 physical map suggests the markers occur in the order 19E07, *D2S307*, *D2S72*, *CTLA4*, *D2S105*, whereas the sequence presented here indicates a different order, 19E07, *D2S307*, *CTLA4*, *D2S72*, *D2S105*, *ICOS*. In light of the discrepancy in *D2S307* position, the disease linkage disequilibrium is modulated downstream from *CTLA4* towards *ICOS*, opening the possibility that the regions near *ICOS* are actually responsible for disease linkage to the IDDM12 locus. We demonstrate that within that range exists numerous other potential targets including *ICOS*, HERV-H, ORFs, ESTs, and mouse syntenic sequences. Polymorphic variation within these sequences may lead to unknown physiological consequences. To generate a higher-resolution map diagnostic for genetic diseases, we examined five microsatellite repeats, SARA 31, *CTLA4* 3′ UTR, SARA 1, SARA 43 (*D2S307*), and SARA 47, periodically located at approximately 50 kb intervals between *CD28* and *ICOS* (Fig. 1). In the 13 individuals examined, no single haplotype emerged from analysis of the five markers when they were used in conjunction. Thus, these markers revealed polymorphisms that are not linked in terms of allelic variation patterning. Most notably, the *ICOS* intron 4 SARA 47 microsatellite repeat may prove useful as a polymorphic marker in genetic studies due to extreme variation with at least six different alleles, with no two identical allelic patterning within the random group of 13 individuals shown here. Beyond the microsatellites analyzed, we identified an additional 348 simple repetitive elements in this region, comprising 2.5% of the total sequence; however, the polymorphic state of these sites has yet to be determined. The complete assembly of the costimulatory receptor region sets the foundation for the high-resolution discrimination of numerous elements within this region associated with disease by the use of the representative panel of microsatellite repeats described here or by the future elucidation of SNPs at desired locations. Further analysis of these polymorphisms within patient populations will allow more precise identification of genomic elements that contribute to autoimmune disease predispositions associated with the 2q33

costimulatory receptor gene cluster, including *CD28*, *CTLA4*, *ICOS*, HERV-H, and other elements.

Indirect evidence suggests that endogenous retroviral genes are linked with numerous autoimmune diseases such as systemic lupus erythematosus, multiple sclerosis, type I diabetes, and Sjogren syndrome [22,23]. It has been postulated that transactivation or dysregulation of genes by neighboring endogenous retroviruses may be a mechanism by which autoimmunity is triggered. The 2q33 HERV-H type endogenous retrovirus could contribute to the modulation of *ICOS* in that this entity is located 366 bp downstream from the *ICOS* 3′ UTR in an antisense orientation with respect to *ICOS*, contains two apparently functional LTRs, and is upregulated upon cell activation. Indeed, we demonstrated that an amplification product spanning the retroviral 3′ LTR to *ICOS* 3′ UTR is obtained from a diverse panel of human genomic DNA, suggesting that the presence of this HERV-H adjacent to *ICOS* is common in the population. Further, we show that differential hybridization was detected for genomic regions corresponding to the *ICOS* 3′ UTR and endogenous retrovirus, an indication that differential hybridizing RNA may be derived from either the *ICOS* promoter or the retroviral LTR promoter.

By reporter gene assays it has been shown that certain HERV-H LTRs are transcriptionally active promoters with dependency on MYB [24] and SP1 [25] transcription factors. Analysis of the HERV-H 3′ LTR in the region studied indicates that two MYB binding sites are present in the U3 domain, in addition to a SP1 binding site within a GC/GT box located directly 3′ to the TATA box (data not shown). The presence of these elements suggests that this retroviral LTR is structurally poised for transcription initiation, and was supported by a positive hybridization signal to T-cell RNA samples using strand-specific transcript probes. Indeed, naturally occurring regulatory antisense transcripts have been described for FGF2 [26] and for complement C4 [27]. Like the HERV-H LTR-driven *ICOS* transcripts, C4 antisense transcripts are driven from the LTR promoters of HERV-K endogenous retrovirus incorporated into intron 9 of the C4 gene. Thus, it is possible that biological events such as mutations or gene rearrangements may lead to the alteration of HERV-H LTR promoter activity and produce antisense *ICOS* transcripts and formation of RNAi [32] that could destabilize normal *ICOS* transcription. Of the cytokine panel that is elicited by ICOS receptor signaling in T-cell activation assays, superinduction of the Th-2 cytokine IL-10 seems to the most prominent [3], although IL-4, IL-5, γ-IFN, TNF-α, and GM-CSF were also produced. Given the complex nature of this cytokine panel, alteration of ICOS signaling could potentiate cytokine skewing and deviate T-cell function.

Recently, *Icos*-deficient mice were shown to be more susceptible to induction of EAE [29], suggesting a role for ICOS in the disease process of EAE. In a subsequent report using SJL mice, administration of blocking anti-ICOS antibody during the antigen priming phase of EAE resulted in exacerbation of disease [30]. Administration of anti-ICOS antibody during the efferent response at a later phase, but before onset of disease, resulted in protection from disease. These results demonstrate that the temporal downregulation of ICOS signaling has an important role in the onset of EAE in mice, and leads us to suggest that such a situation may also occur in analogous settings of human autoimmune disease. In human multiple sclerosis, endogenous retroviral activation has been associated with this disease, but the mechanism for disease onset has not been established [31]. We provide a theoretical framework where exogenous triggering of 2q33 HERV-H LTRs could contribute to the etiology of multiple sclerosis and other autoimmune diseases: production of antisense *ICOS* transcripts during the initial T-cell activation/antigen priming phase could cause the initial repression of ICOS protein synthesis. Priming of autoreactive T cells may then be more likely to result in skewing toward a pathogenic TH1 phenotype.

Other mechanisms also have been proposed by which endogenous retroviruses could lead to immune dysregulation. Retroviral *cis*-regulatory elements may influence transcription of cellular genes involved in immune function, whereas retroviral transactivating elements, like tax in human T-lymphotrophic virus type I or tat in human immunodeficiency virus-I, may induce cellular genes not necessarily proximal to the retroviral element. Indeed, recent evidence shows a direct correlation between HIV serum titer levels and detectable lymphocyte ICOS surface proteins during the time course of infection [32,33]. The mechanism by which HIV infection and ICOS induction are linked remains to be determined. Endogenous retroviruses may also code immunologically active proteins such as the superantigen activity in the LTR of mouse mammary tumor viruses and the nonspecific immunosuppressive activity in mammalian type C retrovirus env protein. Studies of the biological activity of this endogenous retrovirus in relation to *ICOS* transcriptional stability may allow greater insight into whether cytokine balance in immunological disease is linked to this specific region of DNA.

Materials and Methods

BAC clone selection. BAC clones were selected on the basis of positive hybridization to *CTLA4*, *CD28*, or *ICOS* coding sequences (Genome Systems, St. Louis, MO). BAC clone DNA was prepared using Concert Mega Preps BAC protocol followed by restriction endonuclease digestion of 1 μg per sample. Digested samples were electrophoresed in 7% TBE agarose gels followed by electrotransfer onto hybond membranes. Hybridization was performed against random-primed *CTLA4*, *CD28*, or *ICOS* cDNA probes using 0.4% White Rain Shampoo with Conditioner (Gillette, Boston, MA) at 55°C for 1 h followed by washing with 1× SSC, 1% SDS, and then 0.1× SSC, 1% SDS at 55°C until acceptable background was achieved [34].

BAC clone sequencing. BAC clones were shotgun cloned into pUC18 vectors followed by high-throughput sequencing (Lark Technologies, Houston, TX). Briefly, BAC clones were sheared by spray nebulization followed by agarose fractionation and purification of 2–4 kb and 1–2 kb fragments. Fragments were blunt-end cloned into pUC18 *Sma*I site and subsequently used to generate BAC subclone libraries. Contig assembly was initially performed with GAP4 [35] and subsequent manual editing carried out using Sequencher (Gene Codes, Ann

Arbor, MI). Contig gap closure was performed by primer walk sequencing directly on BAC clones using ABI PRISM Big Dye terminator cycle sequencing chemistry and ABI PRISM 373a sequencer. Final assembly and sequence comparison was performed by alignment with GenBank sequences AC010138, AC009965, AF225899, and AF225900. BaC clones described in this study have been deposited in GenBank as accession numbers AF411057, AF411058, and AF41105.

Sequence verification. We verified 2q33 sequence assembly by *Bam*HI, *Eco*RI, and *Hin*dIII digests of BAC clones 22607, 22608, and 22700 and comparison with predicted restriction digest banding patterns. Although fragments from 28,000 kb to 7 bp were generated, only those ranging from greater than 2 kb to less than 12 kb in size were fractionated sufficiently on 0.7% agarose gels for visual analysis. The only notable discrepancy was found by the presence of a 7.7-kb *Bam*HI restriction fragment in BAC clone 22608 not predicted by sequence data, suggesting a base-miscall leading to the elimination of a *Bam*HI site. The sequence results of BAC clone 22700 were further confirmed by restriction mapping the BAC clone using end-labeled oligonucleotide probes as hybridization probes corresponding to predicted *Eco*RI or *Sac*I fragments. Blots were exposed to phosphoimage plates and processed using Fujix image plate reader and Image Reader software. We carried out 29 blot hybridizations with complete accuracy to predicted DNA fragments within BAC 22700. As an external verification of contig assembly, dot-plot analysis (30 bp window, 90% identity) was performed aligning 2q33 sequence with Celera Genomic Axis GA_X8WHR7H (Release 25, Celera Genomics, Rockville, MD 20850). Resultant alignment demonstrated colinearity between the two sequences across 300,000 bp suggesting the correct contig ordering of this genomic region (data not shown).

Sequence analysis. GCG Wisconsin package 10.0 (GCG, Madison, WI) was used for BLAST and FastA database searching. Contigs generated by sequencing were compared to protein databases using TBLASTN to identify potential coding sequences. After final assembly into one contig, sequences were parsed and BLAST searches were performed against GenBank EST and STS databases. Positive EST hits with 80% greater were further BLASTed against GenBank to determine whether cDNA, Unigene, or protein identity could be determined. Complex repeats and ORF prediction was performed by GRAIL (Genomix, Oak Ridge, TN) and DiCTion (Genetics Institute, Cambridge, MA) under default settings. Alignment of *ICOS* genomic sequences was performed with GAP with a gap length penalty set to zero. The alignment output was displayed positionally using PlotSimilarity with an analysis window of 100 nt. Dot plot of mouse and human *ICOS* genomic sequences was performed using GeneWorks (Oxford Molecular Group, Campbell, CA) using a window size of 20 nt and 70% sequence identity cutoff. Cross-species genomic sequence alignment was performed using SIM4 [36] with an F value = 1.3 and word size = 15. Mouse contigs with homologies greater than 35 nt in length were used in further analysis.

Genomic microarray expression analysis. Plasmid preparations of 864 randomly picked colonies from the BAC 22700 subclone library were used as templates for PCR amplification. PCR amplifications were carried out using modified M13 primers in 100 ml reactions containing 10 mM Tris, 1.5 mM $MgCl_2$, 50 mM KCl, 200 mM each dNTP, 200 nM each primer, and 1 unit *Taq* polymerase (Roche Molecular Biochemicals, Mannheim, Germany). PCR products were analyzed by agarose gel electrophoresis and scored for the presence of a single band resulting in 620/864 subclones yielding a robust single band. PCR products were purified using Millipore MultiScreen-FB filter plates essentially as described by the manufacturer (Millipore, Bedford, MA). Dried PCR products were resuspended in 5 M sodium thiocyanate and spotted in duplicate onto Type VI slides (Molecular Dynamics, Sunnyvale, CA) using a GenII arrayer (Molecular Dynamics, Sunnyvale, CA). Probes were prepared by including Cy3 or Cy5 labeled dCTP (Amersham Pharmacia Biotech, Piscataway, NJ) in oligo-(dT) primed first-strand cDNA synthesis reactions from 10 mg total RNA essentially as described [37]. Hybridizations were carried out at 42°C for 16 h in buffer containing 50% formamide, 5× SSC, 0.1% SDS, and 100 mg/ml human COT-1 DNA (Life Technologies, Rockville, MD). The arrays were washed at room temperature once in 1× SSC, 0.2% SDS for 5 min, and twice in 0.1× SSC, 0.2% SDS for 10 min then rinsed in water and dried with compressed nitrogen. Scanning was carried out using a ScanArray 5000 confocal laser scanner (GSI Lumonics, Waltham, MA) and quantitated using ArrayVision 4.0 (Imaging Research, Inc, St. Catharines, ON, Canada). Data from replicate spots on three arrays were combined by taking the average of the log transformed ratio. Differential upregulation was defined as 1.5-fold induction in at least 5 of 6 measurements and having a total signal intensity above a background threshold (1000 for Cy3 + Cy5 on BAC37 reference control.)

Microsatellelite polymorphism analysis. Human donor placental and peripheral blood DNA were used as amplification templates. Single members of oligonucleotide pairs were end-labeled with γ-^{32}P-ATP using T4 polynucleotide kinase (New England Biolabs, Beverly, MA) followed by purification through G25 spin columns. PCR reactions (15 µl) were carried out using Platinum *Taq* (Life Technologies) according to the manufacturer's protocol using 5 pM of each primer and cycled 30 times with the parameters: 95°C for 1 min, 60°C for 1 min, and 72°C for 1 min. Amplified microsatellite DNA was fractionated on Novex QuickPoint Sequencing gels (Invitrogen, Carlsbad, CA). Microsatellite amplification primer pairs used and predicted approximate PCR product size are as follows: SARA 1, 5′-CATGCGGGTTAATACTTAAT-3′, 5′-TTCTCTAGAGGGACAGAACG-3′ (105 bp); SARA 31, 5′-TGCACTCCAGCCTGAGCGAC-3′, 5′-TTCAACACTTAAGAATGGGG-3′ (100 bp); SARA 43, 5′-TATTTCTCCTCTTTCACTGG-3′, 5′-TGACCTGAAATAAACATAGA-3′ (86 bp); SARA 47, 5′-GGTGTTGAAGCATAAAGATG-3′, 5′-TCCCCTCTCCATTGCCTTTC-3′ (104 bp); *CTLA4* 3′ UTR, 5′-TAGCCAGTGATGCTAAAGGTTG-3′, 5′-AACATACGTGGCTCTATGCACA-3′ (106 bp); *ICOS* 3′ UTR retrovirus, 5′-GCAAAGAATA AACATTTGATATTCAGC-3′, 5′-CCCCCCTTTGAATGTAATTTTCCTTTACG-3′ (546 bp).

Acknowledgments

We thank Jim Freeman for assistance in computational analysis of 2q33 sequences and Julie Brown for technical assistance related to GMEA (both from the Genetics Institute at Wyeth Research).

RECEIVED FOR PUBLICATION MARCH 19;
ACCEPTED SEPTEMBER 19, 2001.

References

1. Abbas, A. K., and Sharpe, A. H. (1999). T-cell stimulation: an abundance of B7s. *Nat. Med.* 5: 1345–1346.
2. Balzano, C., Buonavista, N., Rouvier, E., and Golstein, P. (1992). CTLA-4 and CD28: similar proteins, neighbouring genes. *Int. J. Cancer Suppl.* 7: 28–32.
3. Hutloff, A., *et al.* (1999). ICOS is an inducible T-cell co-stimulator structurally and functionally related to CD28. *Nature* 397: 263–266.
4. Ling, V., *et al.* (2000). Cutting edge: identification of GL50, a novel B7-like protein that functionally binds to ICOS receptor. *J. Immunol.* 164: 1653–1657.
5. Yoshinaga, S. K., *et al.* (1999). T-cell co-stimulation through B7RP-1 and ICOS. *Nature* 402: 827–832.
6. Swallow, M. M., Wallin, J. J., and Sha, W. C. (1999). B7h, a novel costimulatory homolog of B7.1 and B7.2, is induced by TNFα. *Immunity* 11: 423–432.
7. Ling, V., *et al.* (1999). Complete sequence determination of the mouse and human CTLA4 gene loci: cross-species DNA sequence similarity beyond exon borders. *Genomics* 60: 341–355.
8. Aicher, A., *et al.* (2000). Characterization of human inducible costimulator ligand expression and function. *J. Immunol.* 164: 4689–4696.
9. Frazer, K. A., *et al* (1997). Computational and biological analysis of 680 kb of DNA sequence from the human 5q31 cytokine gene cluster region. *Genome Res.* 7: 495–512.
10. Hattori, M., *et al.* (2000). The DNA sequence of human chromosome 21. The chromosome 21 mapping and sequencing consortium. *Nature* 405: 311–319.
11. Deng, Z., *et al.* (2000b). Familial primary pulmonary hypertension (Gene PPH1) is caused by mutations in the bone morphogenetic protein receptor-II gene. *Am. J. Hum. Genet.* 67: 737–744.
12. Lee, K. P., *et al.* (1990). The genomic organization of the CD28 gene. Implications for the regulation of CD28 mRNA expression and heterogeneity. *J. Immunol.* 145: 344–352.
13. Shinohara, T., Taniwaki, M., Ishida, Y., Kawaichi, M., and Honjo, T. (1994). Structure and chromosomal localization of the human PD-1 gene (PDCD1). *Genomics* 23: 704–706.
14. Oshima, R. G., Baribault, H., and Caulin, C. (1996). Oncogenic regulation and function of keratins 8 and 18. *Cancer Metastasis Rev.* 15: 445–471.
15. Philpott, A., Krude, T., and Laskey, R. A. (2000). Nuclear chaperones. *Semin. Cell. Dev. Biol.* 11: 7–14.
16. Kim, J. W., *et al.* (1997). Cloning of the human cDNA sequence encoding the NADH:ubiquinone oxidoreductase MLRQ subunit. *Biochem. Mol. Biol. Int.* 43: 669–675.
17. Murphy, P. M. (1993). Molecular mimicry and the generation of host defense protein diversity. *Cell* 72: 823–826.
18. Lacy, D. A., *et al.* (2000). Faithful expression of the human 5q31 cytokine cluster in transgenic mice. *J. Immunol.* 164: 4569–4574.

doi:10.1006/geno.2001.6655, avail[illegible]line at http://www.idealibrary.com on IDEAL

19. Shoemaker, D. D., *et al.* (2001). Experimental annotation of the human genome using microarray technology. *Nature* 409: 922–927.
20. Deng, Z., *et al.* (2000a). Fine mapping of PPH1, a gene for familial primary pulmonary hypertension, to a 3-cM region on chromosome 2q33. *Am. J. Respir. Crit. Care Med.* 161: 1055–1059.
21. Marron, M. P., *et al.* (2000). Genetic and physical mapping of a type 1 diabetes susceptibility gene (IDDM12) to a 100-kb phagemid artificial chromosome clone containing D2S72-CTLA4-D2S105 on chromosome 2q33. *Diabetes* 49: 492–499.
22. Mason, A. L., Xu, L., Guo, L., and Garry, R. F. (1999). Retroviruses in autoimmune liver disease: genetic or environmental agents? *Arch. Immunol. Ther. Exp. (Warsz)* 47: 289–297.
23. Nakagawa, K., and Harrison, L. C. (1996). The potential roles of endogenous retroviruses in autoimmunity. *Immunol. Rev.* 152: 193–236.
24. de Parseval, N., Alkabbani, H., and Heidmann, T. (1999). The long terminal repeats of the HERV-H human endogenous retrovirus contain binding sites for transcriptional regulation by the Myb protein. *J. Gen. Virol.* 80: 841–845.
25. Sjottem, E., Anderssen, S., and Johansen, T. (1996). The promoter activity of long terminal repeats of the HERV-H family of human retrovirus-like elements is critically dependent on Sp1 family proteins interacting with a GC/GT box located immediately 3′ to the TATA box. *J. Virol.* 70: 188–198.
26. Li, A. W., and Murphy, P. R. (2000). Expression of alternatively spliced FGF-2 antisense RNA transcripts in the central nervous system: regulation of FGF-2 mRNA translation. *Mol. Cell. Endocrinol.* 162: 69–78.
27. Schneider, P. M., Witzel-Schlomp, K., Rittner, C., and Zhang, L. (2001). The endogenous retroviral insertion in the human complement C4 gene modulates the expression of homologous genes by antisense inhibition. *Immunogenetics* 53: 1–9.
28. Zamore, P. D., Tuschl, T., Sharp, P. A., and Bartel, D. P. (2000). RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* 101: 25–33.
29. Dong, C., *et al.* (2001). ICOS co-stimulatory receptor is essential for T-cell activation and function. *Nature* 409: 97–101.
30. Rottman, J. B., *et al.* (2001). The costimulatory molecule ICOS plays an important role in the immunopathogenesis of EAE. *Nat. Immunol.* 2: 605–611.
31. Perron, H., and Seigneurin, J. M. (1999). Human retroviral sequences associated with extracellular particles in autoimmune diseases: epiphenomenon or possible role in aetiopathogenesis? *Microbes Infect.* 1: 309–322.
32. Buonfiglio, D., *et al.* (2000). The T cell activation molecule H4 and CD28-like molecule ICOS are identical. *Eur. J. Immunol.* 30: 3463–3467.
33. Lucia, M. B., *et al.* (2000). Expression of the novel T cell activation molecule hpH4 in HIV-infected patients: Correlation with disease status. *AIDS Res. Human Retroviruses* 16: 549–557.
34. May, B. P. (1998). Southern hybridization in shampoo. *Biotechniques* 25: 582.
35. Bonfield, J. K., Rada, C., and Staden, R. (1998). Automated detection of point mutations using fluorescent sequence trace subtraction. *Nucleic Acids Res.* 26: 3404–3409.
36. Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M., and Miller, W. (1998). A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* 8: 967–974.
37. Schena, M., *et al.* (1996). Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci. USA* 93: 10614–10619.